JMS Journées de méthodologie statistique de l'Insee 2018

RECONSTITUTION DES MOUVEMENTS DE MAIN-D'ŒUVRE A PARTIR DE LA DECLARATION SOCIALE NOMINATIVE

QUELLE CORRECTION APPORTEE DANS LA PHASE DE MONTEE EN CHARGE DU DISPOSITIF ?

Kévin MILIN (*)

(*) Dares, département emploi

kevin.milin@travail.gouv.fr

Mots-clés : déclaration sociale nominative, données administratives, mouvements de main-d'œuvre, redressement de la non-réponse totale

Résumé

La déclaration sociale nominative (DSN) a été progressivement mise en place. Elle s'établit dans un premier temps sur la base du volontariat, dès 2013. Ensuite, après quelques obligations intermédiaires, elle s'est généralisée à tous les établissements relevant du régime général début 2017. Ce nouveau dispositif a pour but de remplacer un grand nombre de déclarations pour les employeurs, notamment les déclarations de mouvements de main-d'œuvre (DMMO).

Auparavant construites à partir de la DMMO et de son enquête associée (EMMO), les statistiques de mouvements de main-d'œuvre (MMO) sont désormais reconstituées à partir des DSN. Ce changement de source a nécessité un travail méthodologique conséquent. L'une des étapes consiste à corriger les données de la « non-réponse totale », en affectant un poids à chaque établissement ayant déposé des DSN. Cette correction permet de pallier la non-réponse des établissements réputés concernés par cette déclaration et de garantir la représentativité des données finales. La méthode retenue se base sur les probabilités estimées de déclaration des établissements. Ces estimations sont issues de modélisations qui prennent en compte les non-linéarités, par exemple dans la relation entre l'effectif d'un établissement et sa propension à faire une DSN, en transformant les variables quantitatives en variables qualitatives. Le nombre de catégories, ainsi que leurs contours, ne sont pas les mêmes selon les secteurs et évoluent dans le temps : les spécifications sont adaptées à chaque phase de montée en charge de la DSN. Enfin, l'utilisation de la déclaration préalable à l'embauche (DPAE), comme variable auxiliaire fortement corrélée aux variables d'intérêt des MMO, permet de confirmer que les données finales sont correctement corrigées des biais de non-déclaration des établissements, notamment lors de la montée en charge du dispositif des DSN.

Introduction

Les mouvements de main-d'œuvre (MMO) recensent l'ensemble des embauches et des fins de contrats de travail au niveau des établissements. Ils permettent de mesurer les entrées et les sorties selon le type de contrat (contrat à durée déterminée – CDD/contrat à durée indéterminée – CDI), la durée des contrats, les motifs de rupture, mais aussi selon les caractéristiques des salariés et des établissements. Ils sont traditionnellement rapportés à l'effectif moyen des établissements.

Jusqu'en 2015, ces statistiques étaient élaborées à partir de deux sources : une déclaration mensuelle obligatoire pour les établissements de plus de 50 salariés (DMMO), et une enquête trimestrielle pour les établissements de moins de 50 salariés (EMMO). Dans chacune de ces deux sources les établissements déclaraient directement des flux, c'est-à-dire uniquement leurs embauches et leurs fins de contrats.

À partir de 2013, la déclaration sociale nominative (DSN) a été progressivement mise en place. Elle est constituée de l'ensemble des paies versées par les établissements chaque mois. Elle vise à remplacer un grand nombre de déclarations administratives réalisées par les entreprises, dont celles portant sur les mouvements de main-d'œuvre. Dans ce contexte, au cours de l'année 2015, le taux de réponse à la DMMO/EMMO a nettement diminué car de plus en plus d'établissements sont entrés dans le dispositif de la DSN. C'est la raison pour laquelle la publication des données sur les MMO a été suspendue à partir du deuxième trimestre.

Ainsi, un travail conséquent a été mené au sein de la Dares, afin de reconstituer les mouvements de main-d'œuvre à partir des DSN. Cinq principales étapes jalonnent ce travail :

a. Reconstituer des flux de main-d'œuvre à partir des stocks de contrats (les DSN sont issues de déclarations de salaires). L'objectif est de suivre, de déclaration en déclaration, la vie des contrats afin de détecter les embauches et les fins de contrat. Ce travail est géré par un système d'information (cf. Barlet et *al.*, 2015).
b. Construire des référentiels d'établissements reconstituant le champ historique des MMO, c'est-à-dire le champ privé hors agriculture de France métropolitaine, avec une prise en compte partielle de la démographie des établissements. Ceci donne une population dont les DSN sont attendues pour la construction des données.
c. S'assurer que les établissements qui ont remis des DSN soient représentatifs de l'ensemble du champ MMO. Pour cela, un poids est affecté aux établissements déclarants afin de pallier la non-déclaration des établissements réputés concerné par la DSN.
d. Corriger les problèmes de qualité dans les DSN. En particulier, il parait essentiel d'affecter un motif de fin aux contrats qui ont disparu d'une déclaration à l'autre (*i.e.* aux fins de contrat qui n'ont pas été correctement déclarées).
e. Corriger les données historiques des biais de sous-déclaration. Une rétropolation au niveau des données détaillées a été mise en place.

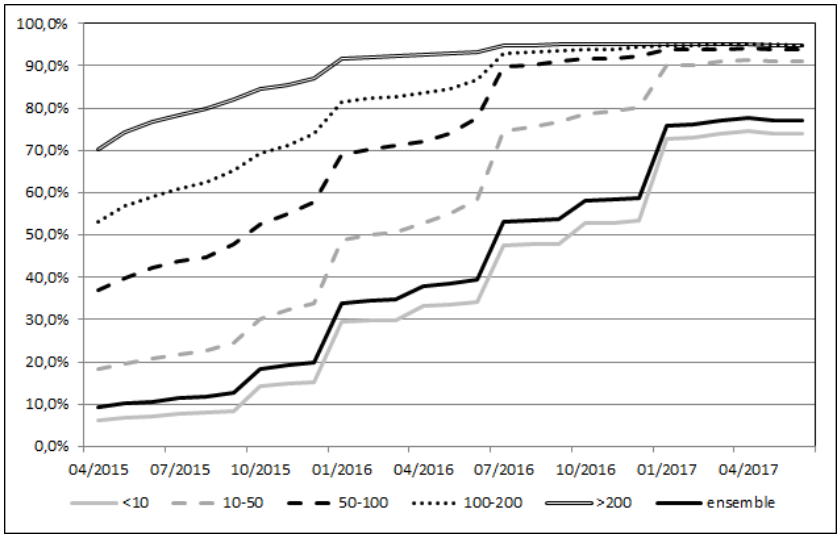
Ces étapes sont en grande partie décrites dans un document d'études de la Dares, paru en juin 2018 et intitulé « Reconstitution des mouvements de main-d'œuvre depuis 1993 : guide méthodologique » (Milin, 2018). Cette présente contribution aux Journées de méthodologie statistique 2018 en est essentiellement un extrait. Elle décrit plus spécialement la modélisation retenue pour estimer les probabilités des établissements à déposer des DSN. Ces probabilités sont nécessaires pour redresser les données de la non-déclaration et garantir la représentativité des statistiques finales. Elle montre également que la qualité de la correction mise en place dépend en partie de la prise en compte des non-linéarités, par exemple dans la modélisation, dans la relation entre l'effectif d'un établissement et sa propension à déposer une DSN.

1. Une montée en charge progressive de la déclaration sociale nominative

Historiquement, les statistiques de mouvements de main-d'œuvre (MMO) couvrent les établissements de France métropolitaine, relevant du champ privé hors agriculture et hors intérimaires. Elles ne prennent en compte que partiellement les effets démographiques, liés aux changements de statut des établissements (de non-employeur à employeur) et à leur cessation d'activité. De fait, le dispositif MMO nécessite de faire appel au référentiel Sirus (Système d'identification au répertoire des unités statistiques), qui fournit des informations sur les établissements avec un certain délai. Dans ce contexte, les MMO portent majoritairement sur les établissements employeurs depuis plus de deux ans. Sur le champ des établissements de plus de 50 salariés soumis à la DMMO, ceux réalisant pour la première fois une déclaration sont toutefois pris en compte dès l'année suivante.

Ce champ a été reconstitué sur les données issues de la DSN, dans une base dite de référence, élaborée pour chacune des années 2015 à 2017. Ces bases comprennent tous les établissements réputés concernés par les MMO, suivant la définition historique rappelée ci-dessus.

Au moyen de ces bases de référence, il est possible de calculer le taux de déclaration sur le champ MMO. Considéré globalement, il s'établit, à environ 10 % en avril 2015, et s'élève à plus de 75 % en 2017. La montée en charge du dispositif DSN s'est faite par paliers, conformément à la législation[1].



Graphique 1 – Taux de réponse à la DSN parmi les établissements concernés, selon la taille des établissements

2. Traitement de la non-déclaration des établissements

En 2015 et 2016, la DSN étant dans une phase de déploiement, tous les établissements n'en ont pas effectuées. En 2017, le taux de réponse à la DSN n'atteint pas encore les 100 % (cf. paragraphe 1). Une méthode de redressement de la non-réponse totale permet de corriger les effets de cette non-exhaustivité, en affectant un poids à chaque établissement déclarant pour garantir la représentativité des statistiques produites sur les mouvements de main-d'œuvre.

Plus précisément, plusieurs étapes jalonnent ce travail :

a. Les probabilités de déclaration de chaque établissement sont estimées pour chaque mois. Il convient de prendre en compte les effets de non-linéarités dans les estimations. Par

[1] Pour plus de précisions sur les différents paliers, voir le document d'études de la Dares (Milin, 2018).

exemple, la probabilité d'entrer dans le dispositif des DSN n'augmente pas linéairement en fonction de l'effectif de l'établissement.

b. Des groupes de réponse homogène (GRH) sont ensuite créés à partir des probabilités estimées de déclaration. Le poids affecté à chaque établissement vaut alors l'inverse du taux de réponse dans le GRH associé.
c. Un calage sur marge est enfin appliqué, notamment pour diminuer la variance des estimateurs.

Dans la suite du document, le principe général de la correction est tout d'abord explicité. Puis, la modélisation retenue pour estimer les probabilités de déclaration des établissements y est décrite. Enfin, la qualité de la correction apportée est jugée à l'aide des déclarations préalables à l'embauche (DPAE) de l'Acoss (Agence centrale des organismes de sécurité sociale). La méthode de construction des GRH et le calage sur marges ne sont pas présentés ici.

2.1. Notations et présentation de la méthode de redressement de la non-déclaration

Tout au long de ce document, on note *y* une variable d'intérêt présente dans les données issues des DSN, comme par exemple le nombre d'entrées ou de sorties sur une période considérée. Par ailleurs, ayant à disposition un référentiel qui liste les établissements pour lesquels une DSN est attendue, on note r_i la variable indicatrice de réponse pour l'unité *i* : $r_i = 1$ si l'établissement *i* a effectué une DSN, et $r_i = 0$ dans le cas contraire. Notons *bdr* la base de référence, et *R* l'ensemble des déclarants (ou répondants), c'est-à-dire les établissements pour lesquels $r_i = 1$.

La méthode de correction proposée consiste à expliquer le mécanisme de réponse en attachant à chaque établissement une probabilité de répondre (de remettre une DSN) qui varie en fonction de ses caractéristiques, et de pondérer chaque établissement répondant par l'inverse de sa probabilité de réponse estimée.

Notons $p_i = P(r_i = 1),\ i \in bdr$ la probabilité que $i \in bdr$ ait remis une DSN. Cherchons à estimer le total de la variable d'intérêt (*i.e.* le nombre total d'entrées ou de sorties sur la période). En l'absence de non-réponse, ce total s'écrit :

$$t_y = \sum_{i \in bdr} y_i$$

Comme la base de référence est exhaustive, ce total est connu de façon exacte.

La correction de non-réponse consiste à remplacer l'expression précédente par une somme sur l'ensemble des répondants, avec prise en compte de la pondération :

$$\widetilde{t_y} = \sum_{i \in R} \frac{y_i}{p_i}$$

Étant donné que les probabilités de réponse ne sont pas connues, elles sont estimées *via* un modèle : $p_i = f(\boldsymbol{x}_i, \theta)$, où $\boldsymbol{x}_i$ représente l'information auxiliaire disponible dans la base de référence, et où θ est un vecteur de paramètres inconnus à estimer. En notant $\hat{p}_i = f(\boldsymbol{x}_i, \hat{\theta})$, la probabilité estimée de réponse de l'établissement *i,* l'estimateur du total de la variable d'intérêt *y* est donné par :

$$\widehat{t_y} = \sum_{i \in R} \frac{y_i}{\hat{p}_i}$$

La qualité de l'estimation est fortement liée à la qualité du modèle d'explication de la non-réponse[2]. De plus, il faut prendre garde à la dispersion éventuelle des probabilités estimées, qui peut entrainer une variance importante de l'estimateur $\widehat{t_y}$ (liée à de petites probabilités de réponse).

[2] Si le modèle de non-réponse n'est pas correctement spécifié, l'estimateur qui en découle peut être considérablement biaisé.

Afin d'apporter de la robustesse, on a recours à des classes de pondération, plutôt que d'utiliser directement les probabilités de déclaration (cf. Ardilly, 2006). On divise la base de référence en *T* classes (C_1, C_2, … C_T), qui comportent à la fois des déclarants et des non-déclarants, et on affecte à chaque déclarant un poids égal à l'inverse du taux de réponse observé dans sa classe. En adoptant cette stratégie, l'estimateur prend la forme suivante :

$$\ddot{t_y} = \sum_{i \in R} \frac{y_i}{\widehat{p_i}} = \sum_{c=1}^{T} \sum_{i \in C_c} \frac{r_i}{\widehat{p_c}} y_i = \sum_{c=1}^{T} N_c \overline{y_{rc}}$$

où $N_c = \sum_{i \in C_c} 1$; $\widehat{p_c} = \frac{\sum_{i \in C_c} r_i}{N_c}$; $\overline{y_{rc}} = \frac{\sum_{i \in C_c} r_i y_i}{\sum_{i \in C_c} r_i}$

Et le biais de non-déclaration associé à l'estimateur du total de *y* prend la forme :

$$B(\ddot{t_y}) = E(\ddot{t_y} - t_y) \cong \sum_{c=1}^{T} \frac{1}{\bar{p}_c} \sum_{i \in C_c} (p_i - \overline{p_c})(y_i - \overline{y_c})$$

avec : $\bar{p}_c = \frac{\sum_{i \in C_c} p_i}{N_c}$ et $\bar{y}_c = \frac{\sum_{i \in C_c} y_i}{N_c}$

Cette écriture suggère que le biais dû à la non-déclaration est petit :

- si la probabilité moyenne de déclaration dans chaque classe est grande (*i.e.* que le taux de réponse dans la classe est grand) ;
- ou si la covariance entre la probabilité de faire une déclaration et la variable d'intérêt est faible.

Ainsi, il parait naturel de former des classes homogènes par rapport aux probabilités de réponse, puisque les variables d'intérêt sont nombreuses.

Plusieurs méthodes de formation des classes ont été testées sur le mois d'octobre 2015. Une méthode dite des scores s'est avérée la plus efficace, et a donc été retenue. Cette dernière se divise en trois étapes :

a. Modélisation du score, c'est-à-dire estimation de la probabilité de déposer une DSN pour chaque établissement apparaissant dans le référentiel ;
b. Formation des classes en se basant sur les probabilités de déclaration estimées selon la méthode décrite par Haziza et Beaumont (2007) ;
c. Affectation d'un poids à chaque établissement (inverse du taux de réponse de la classe).

Comme la construction des classes se base principalement sur les probabilités estimées de déclaration, et donc sur le modèle utilisé, l'étape d'élaboration de modèles de déclaration/non-déclaration s'avère essentielle dans le processus d'affectation de poids aux établissements ayant fait des DSN. Les parties suivantes détaillent la manière de construire des modèles de non-réponse en DSN, ainsi que les éléments qui permettent leur validation.

2.2. Prise en compte des non-linéarités

Différentes variables sont intégrables dans la modélisation des probabilités de déclaration : le secteur d'activité, l'effectif salarié ou la zone géographique. Les variables retenues doivent caractériser les établissements et doivent être corrélées à leur comportement de déclaration. Par exemple, la taille des établissements ou leur appartenance à un groupe peut indirectement capter l'existence d'une structure interne, comme un service de ressources humaines ou de comptabilité, en charge des déclarations légales.

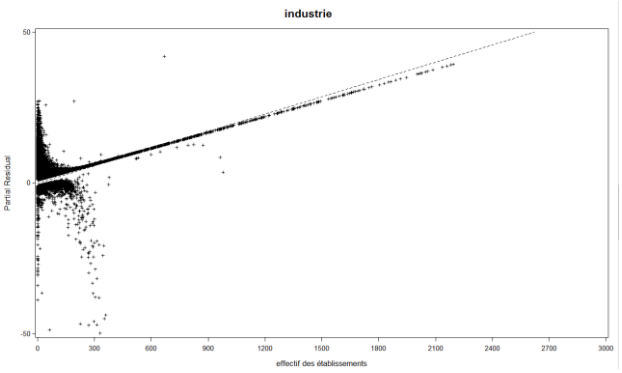
Un modèle paramétrique, tel que présenté en encadré, intègre deux variables quantitatives : l'effectif de l'établissement, ainsi que le nombre d'établissements de l'entreprise à laquelle appartient l'établissement considéré. Il est difficile d'imaginer que le lien entre la propension à

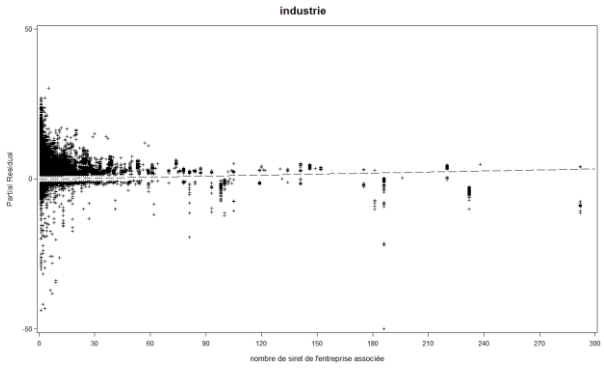
effectuer une DSN et ces deux variables soit totalement linéaire : il peut, par exemple, exister des effets de seuil, c'est-à-dire qu'à partir d'une certaine taille de l'établissement et de l'entreprise, l'effet sur la probabilité de faire une DSN est marginal.

Afin de tester l'existence de non-linéarité, plusieurs méthodes peuvent être utilisées, notamment :

- Un modèle semi-paramétrique (*Generalized Additive Model*) intégrant les mêmes variables que le modèle dit « de base », du type : $logit(p_i) = ln(\frac{p_i}{1-p_i}) = f(\boldsymbol{X}_i) + \boldsymbol{\beta}.\boldsymbol{Z}_i$, avec $\boldsymbol{X}_i$ et $\boldsymbol{Z}_i$ des informations auxiliaires associées à l'établissement *i*, $\boldsymbol{\beta}$ un vecteur de paramètres à estimer, *f* une fonction de forme inconnue. Cette méthode est similaire à celle présentée par Da Silva et Opsomer (2009), et permet dans le même temps d'estimer les probabilités de déclaration pour chaque établissement. Ces modèles ont été testés : ils nécessitent des temps de calcul trop importants.
- L'analyse des résidus partiels[3]. Cela consiste à tracer les résidus partiels du modèle « de base » (cf. encadré), associés aux deux variables quantitatives, en fonction de ces mêmes variables quantitatives. Si le tracé est linéaire alors le modèle initial est accepté. *A contrario*, si une tendance non linéaire se dégage, le modèle initial doit être modifié. Par exemple, il est possible de :
    - remplacer la variable quantitative en question par une fonction de celle-ci donnant la même tendance que celle observée ;
    - ou encore de modifier cette variable en variable catégorielle.

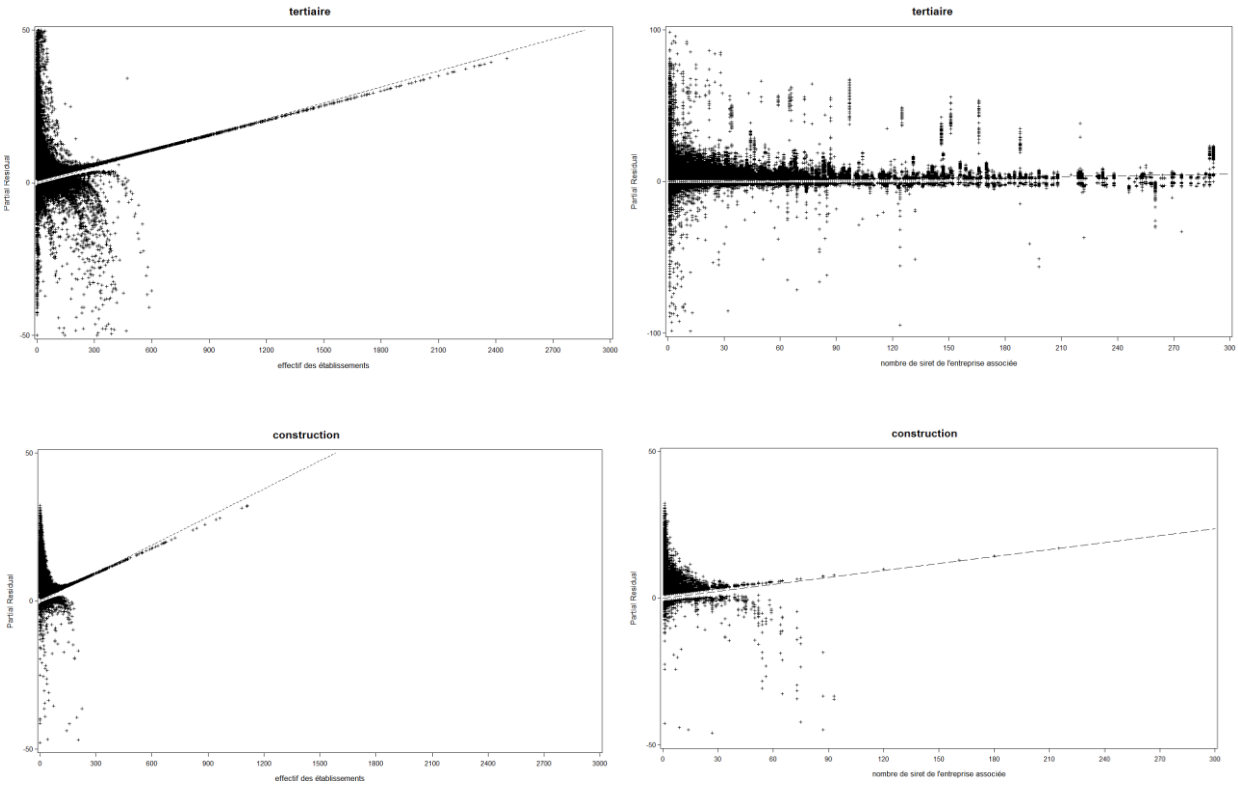
Cette dernière méthode est donc appliquée au modèle « de base » présenté en encadré. Au regard des graphiques 2 à 7, une structure particulière se dégage des résidus partiels associés à l'effectif, et dans une moindre mesure, de ceux associés au nombre d'établissements par entreprise. En conséquence, il convient de prendre en compte cette non-linéarité dans les modèles. Pour cela, la création de classes pour rendre les variables quantitatives qualitatives est préférée à l'utilisation d'une fonction, puisqu'elle permet l'économie de conjectures sur la forme de la fonction la plus adaptée (cf. *infra*).





[3] Les résidus partiels associés à la variable *X* se définissent de la manière suivante :

$$\hat{\varepsilon}_x = \frac{r_{i-}\widehat{p_i}}{(1-\widehat{p_i}).\widehat{p_i}} + \beta_x.X$$



Graphiques 2 à 7 – Résidus partiels calculés à partir du modèle « de base », pour le mois d'octobre 2015.

Encadré – Une première approche pour modéliser les probabilités de déclaration

Un modèle dit « de base » a été mobilisé en première approche, pour estimer les probabilités de déclaration des établissements pour un mois donné, sur chaque grand secteur (industrie, construction, tertiaire). Il s'écrit :

$$logit(p_i) = ln(\frac{p_i}{1-p_i}) = \gamma + \rho.eff_i + \sigma.n_i^{siret} + \sum_{1 \le j \le n_{apet}-1} \alpha_j.1_{apet_i=apet_j} + \sum_{1 \le j \le n_{region}-1} \beta_j.1_{region_i=region_j}$$

Avec :

- eff_i l'effectif de référence de l'établissement *i* (non catégorisé) ;
- n_i^{siret} le nombre d'établissements de l'entreprise à laquelle l'établissement *i* appartient (non catégorisé) ;
- $apet_i$ le secteur d'activité de l'établissement *i*, au niveau le plus fin ; avant octobre 2015, des niveaux plus agrégés sont utilisés afin de garantir la convergence de l'algorithme itératif permettant l'obtention de l'estimateur du maximum de vraisemblance de $(\gamma, \rho, \sigma, \boldsymbol{\alpha}, \boldsymbol{\beta})$;
- $region_i$ la zone géographique d'implantation de l'établissement *i*.

2.3. Recherche de modèle par période

Si les estimations sont effectuées sur les trois grands secteurs d'activité (industrie, construction, tertiaire), des différences subsistent entre les modèles retenus. En effet, pour un secteur donné, les variables sélectionnées dans le modèle ne sont pas forcément les mêmes d'un mois à l'autre, tout comme les classes retenues (nombres et bornes) pour transformer les variables quantitatives en variables qualitatives.

La recherche de modèle n'est néanmoins pas effectuée pour chaque mois, mais selon les paliers qui correspondent aux différentes phases de montée en charge de la DSN (cf. paragraphe 1).

2.4. Variables utilisées

Les modèles retiennent plusieurs variables :

- l'effectif de référence de l'établissement (catégorisé) ;
- le secteur d'activité ;
- la zone géographique d'implantation de l'établissement (région) ;
- le nombre d'établissements de l'entreprise à laquelle l'établissement appartient (catégorisé) ;
- la catégorie juridique de l'entreprise ;
- la date de création de l'établissement (constitution de plusieurs classes[4]) ;
- le chiffre d'affaire de l'entreprise ; la date d'entrée dans le dispositif des DSN dépend des montants passés des cotisations et contributions sociales des entreprises. Le chiffre d'affaire de l'entreprise est alors la variable qui permet d'approcher le plus possible les règles d'entrée dans le processus des DSN.

De même, il est possible de rajouter des variables d'interactions pour augmenter la dimension explicative du modèle. Par exemple, dans le modèle purement additif (modèle de base), l'hypothèse sous-jacente est que, pour une région donnée, les coefficients associés à la variable effectif sont identiques pour les différents secteurs (cf. graphique 8). Dans un modèle avec des interactions (entre effectifs et secteurs), cette hypothèse est relâchée, les pentes peuvent être différentes (cf. graphique 9).

Le nombre de variables d'interactions de premier ordre à introduire est potentiellement important. Cependant, toutes ces variables ne peuvent pas être intégrées au modèle, au risque de nuire à l'estimation des paramètres. Par exemple, un trop grand nombre de coefficients à estimer peut amener :

- à une trop forte complexité calculatoire ;
- à de la séparabilité des données, impliquant la non-convergence de l'estimateur du maximum de vraisemblance ;
- ou encore à un sur-apprentissage.

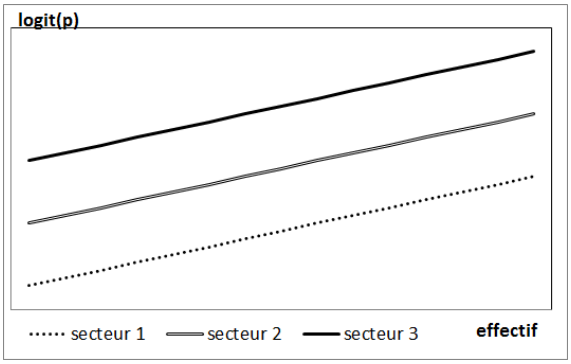
En conséquence, plusieurs variables d'interactions ne sont pas testées dans la modélisation : d'une part celles qui intègrent la catégorie juridique et la région, d'autre part celles qui sont composées de variables dont le V de Cramer est inférieur à 0,15 (les interactions basées sur de fortes dépendances sont donc rejetées). Ainsi, entre cinq et six interactions sont testées, les plus fréquentes étant :

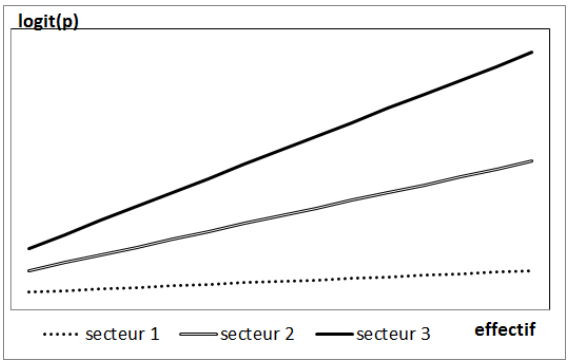
[4] Pour les années de création des établissements, on considère six classes différentes :
- les établissements créés avant 1998 ;
- entre 1998 et 2005 ;
- entre 2006 et 2010 ;
- entre 2011 et N-3, avec N l'année d'intérêt ;
- ceux crées après N-2 ;
- les établissements dont la date de création est inconnue ou incertaine.

- secteur d'activité x effectif ;
- secteur d'activité x chiffre d'affaire de l'entreprise (uniquement pour le secteur industriel) ;
- secteur d'activité x année de création de l'établissement ;
- effectif x nombre d'établissements dans l'entreprise ;
- effectif x année de création de l'établissement ;
- date de création de l'établissement x nombre d'établissements dans l'entreprise ;
- chiffre d'affaire x année de création de l'établissement.

Malgré cette restriction à quelques variables d'interactions, la convergence des estimateurs n'est pas nécessairement assurée. Plusieurs solutions permettent de pallier ce problème :

- passer à des nomenclatures plus agrégées ;
- retirer les interactions qui offrent les plus grands degrés de liberté ;
- retirer les variables qui ne sont pas, *a priori*, significatives (tests basés sur l'estimation issue de la dernière itération de l'algorithme pour l'estimation du maximum de vraisemblance, variables d'interactions comprises).





Graphiques 8 et 9 – Illustrations d'un modèle additif et d'un modèle avec interactions

Ainsi, le modèle s'écrit sous la forme suivante :

$$
\begin{aligned}
&logit(p_i) \\
&= \beta^0 + \beta^1.eff_i + \beta^2.n_i^{siret} \\
&+ \sum_{1\le j\le n_{secteur}-1} \beta_j^3.1_{secteur_{i=}secteur_j} + \sum_{1\le j\le n_{region}-1} \beta_j^4.1_{region_{i=}region_j} \\
&+ \sum_{1\le j\le n_{effc}-1} \beta_j^5.1_{effc_{i=}effc_j} + \sum_{1\le j\le nc^{siret}-1} \beta_j^6.1_{nc_i^{siret}=nc_j^{siret}} + \sum_{1\le j\le n_{cj}-1} \beta_j^7.1_{cj_{i=}cj_j} \\
&+ \sum_{1\le j\le n_{cac}-1} \beta_j^8.1_{cac_{i=}cac_j} + \sum_{1\le j\le n_{age}-1} \beta_j^9.1_{age_{i=}age_j} \\
&+ \sum_{1\le j\le n_{secteur}.n_{age}-2} \beta_j^{10}.1_{secteur_{i=}secteur_j}.1_{age_{i=}age_j} \\
&+ \sum_{1\le j\le nc^{siret}.n_{age}-2} \beta_j^{11}.1_{nc_i^{siret}=nc_j^{siret}}.1_{age_{i=}age_j} \qquad (*)
\end{aligned}
$$

Avec :

- eff_i et $effc_i$ l'effectif et l'effectif catégorisé de l'établissement *i* ;
- n_i^{siret} le nombre d'établissements de l'entreprise dans laquelle l'établissement *i* appartient (nc_i^{siret} étant la même variable, mais catégorisée) ;
- $secteur_i$ le secteur d'activité de l'établissement *i* exprimé dans la nomenclature A88 ;

- $region_i$ la zone géographique d'implantation de l'établissement *i* ;
- age_i l'année de création de l'établissement *i*, en cinq classes (avant 1999, 1999-2005, 2006-2010, 2011-2013, après 2014) ;
- cj_i la catégorie juridique de l'entreprise associée à l'établissement *i* (les sociétés commerciales sont différenciées des autres sociétés) ;
- cac_i le chiffre d'affaire catégorisé de l'entreprise intégrant l'établissement *i.*

2.5. Non-linéarité : méthode retenue pour la création des classes

Au regard des résidus partiels associés à la variable effectif, l'existence d'une non-linéarité dans la relation entre la propension d'un établissement à effectuer une DSN et son effectif est vérifiée. Pour se prémunir de tout effet non linéaire entre variables d'intérêt et variables explicatives, les exogènes quantitatives sont catégorisées (effectif de l'établissement, chiffre d'affaire de l'entreprise, nombre d'établissements dans l'entreprise).

Toutefois, pour catégoriser une variable, le choix du nombre de classes, et de leurs bornes peut s'avérer délicat, et reste généralement arbitraires. En outre, il est souvent retenu des quartiles ou quintiles, sans que cela ne garantisse que ce soit le meilleur choix. Afin que les classes soient les plus pertinentes possibles, la méthodologie suivante est appliquée pour chaque secteur et pour chaque palier :

a. Découpage des distributions des effectifs et des chiffres d'affaire par le moyen d'un jeu de 20 quantiles. Étant donné que le chiffre d'affaire de l'entreprise est une variable continue, ce découpage permet d'obtenir 21 classes de taille égales (plus une pour les entreprises pour lesquelles le chiffre d'affaire n'est pas renseigné) ; au contraire, l'effectif étant une variable discrète, le nombre de classes n'est pas constant ;
b. Estimation du modèle (*), qui comprend toutes les variables préalablement citées, exceptées les variables d'interactions composées de l'effectif de l'établissement et du chiffre d'affaire de l'entreprise. Le retrait de ces quelques variables d'interactions permet notamment de faciliter l'écriture des tests de l'étape 3 ;
c. Application des tests d'égalité suivants[5] :
   $\forall j \in \left[1; n_{effc} - 2\right],\ \beta_j^5 = \beta_{j+1}^5$ et $\forall q \in \left[1\,; n_{cac} - 2\right],\ \beta_q^8 = \beta_{q+1}^8$
d. Fusion de la j^{ième} et (j+1)^{ième} classe d'effectif en lien avec le test qui présente la statistique de test la plus faible (en valeur absolue), et de la q^{ième} et (q+1)^{ième} classe du chiffre d'affaire ;
e. Réitération des étapes b, c et d jusqu'à ce que tous les tests d'égalité soient rejetés au niveau 5 %.

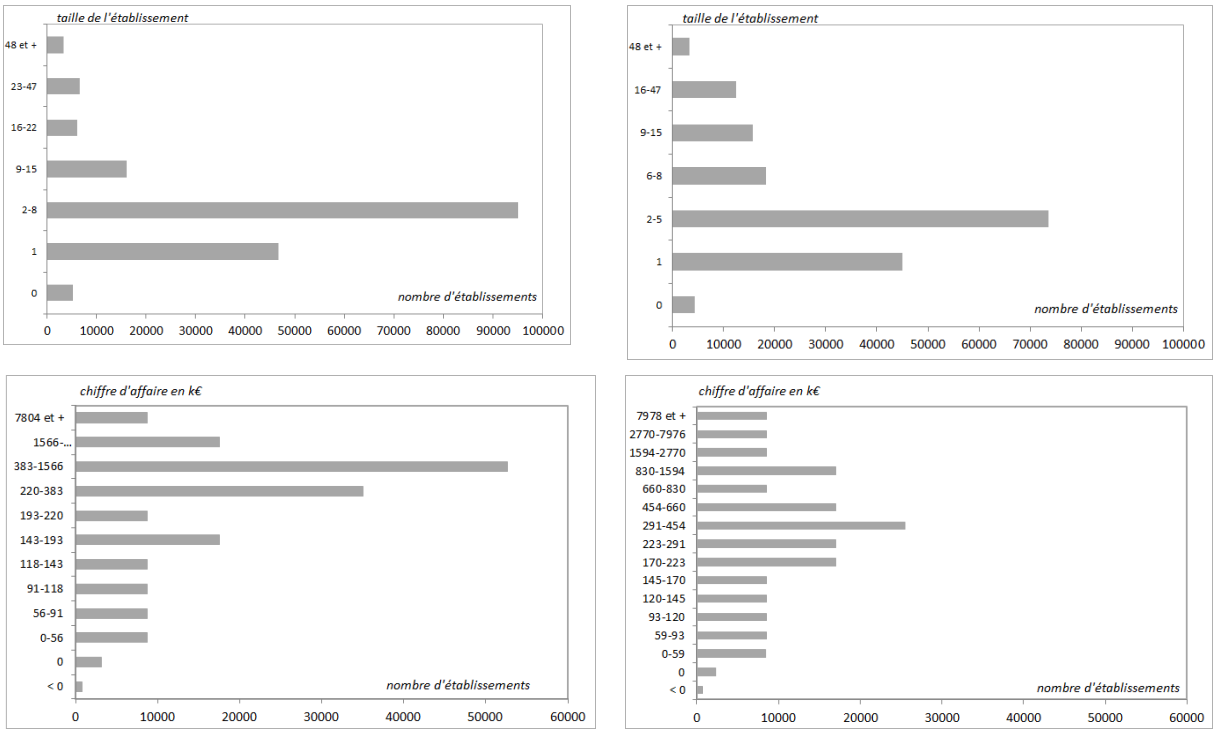
Ainsi, l'application de cet algorithme débouche sur la création de classes pour la variable d'effectif et celle du chiffre d'affaire. Le nombre de classes, ainsi que leurs bornes diffèrent d'un palier à l'autre, et d'un secteur à l'autre (industrie, construction, tertiaire). Par exemple, dans le secteur de la construction, suite à la catégorisation de l'effectif, on obtient le même nombre de classes en janvier et juillet 2016, mais le découpage diffère (cf. graphiques 10 et 11). Par ailleurs, pour le chiffre d'affaire, douze classes ressortent de l'algorithme en janvier 2016, contre seize classes en juillet 2016 (cf. graphiques 12 et 13).

En outre, il est possible de vérifier facilement la présence de non-linéarité avec les modèles qui ont servi à la constitution des classes. Par exemple, en juillet 2016 dans le secteur de la construction :

[5] Pour rendre les tests interprétables, les matrices de variance-covariance sont préalablement modifiées, afin d'adapter les tests aux effets de sur-dispersion (respectivement de sous-dispersion). Si la sur-dispersion (resp. sous-dispersion) n'intervient pas dans l'estimation des paramètres, elle rend les tests trop significatifs (resp. pas assez significatifs).

- Le coefficient associé à l'effectif s'établit à -0,20 pour les établissements ayant seulement 1 salarié. Il s'élève à +0,27 pour les établissements ayant entre 16 et 47 salariés, alors que pour les établissements de plus de 48 salariés, le coefficient associé à la classe d'effectif est nul (cf. tableau 1) ;
- Le paramètre associé aux chiffres d'affaire importants (plus de 7978 k€) est plus bien fort que ceux des chiffres d'affaire plus modestes (1,69 contre un coefficient inférieur à 1,04 pour les autres classes ; cf. tableau 2), conformément aux règles d'entrée dans le dispositif des DSN.

janvier 2016 *juillet 2016*



Graphiques 10 à 13 – Exemples de classes retenues pour les effectifs et les chiffres d'affaires, pour les mois de janvier et juillet 2016, dans le secteur de la construction.

mai-15	effectif	0	1	2-3	4-5	6-10	11-22	23-47	48 et +
	coefficient	-1,77	-0,46	-0,21	0,13	0,47	0,68	0,47	0,00
oct-15	effectif	1	2-7	8-48	49-1516	0,00			
	coefficient	-0,02	0,08	0,18	0,39	0,00			
janv-16	effectif	0	1	2-8	9-15	16-22	23-47	48 et +	
	coefficient	-0,96	0,11	0,17	0,26	0,19	0,07	0,00	
avr-16	effectif	0	1	2-2	3-8	9-22	23-47	48 et +	
	coefficient	-0,66	0,07	0,17	0,12	0,18	0,00	0,00	
juil-16	effectif	0	1	2-5	6-8	9-15	16-47	48 et +	
	coefficient	-0,82	-0,20	-0,07	-0,01	0,15	0,27	0,00	
oct-16	effectif	0	1	2-5	6-8	9-10	11-15	16-47	48 et +
	coefficient	-0,75	-0,31	-0,13	-0,06	0,08	0,15	0,29	0,00
janv-17	effectif	0	1	2-2	3-4	5-10	11-13	14-20	21-43	44 et +
	coefficient	-1,02	-0,63	-0,16	-0,07	0,06	-0,09	0,34	0,55	0,00

Tableau 1 – Estimation des paramètres associés aux classes des effectifs (secteur de la construction)

mai-15	chiffre d'affaire en k€	non-renseigné	< 0	0-58	58-287	287-824	824-2725	2725-7738	7743 et +										
	coefficient	0,00	1,01	-1,73	-0,82	-0,65	-0,26	0,45	2,86										
oct-15	chiffre d'affaire en k€	non-renseigné	< 0	0-61	61-93	93-120	120-223	223-292	292-389	389-2784	2785-7925	7928 et +							
	coefficient	0,00	-0,83	-1,04	-0,27	0,03	0,26	0,44	0,35	0,43	0,59	1,93							
janv-16	chiffre d'affaire en k€	non-renseigné	< 0	0-56	56-91	91-118	118-143	143-193	193-220	220-383	383-1566	1566-7802	7804 et +						
	coefficient	0,00	-1,19	-1,38	-0,38	0,01	0,14	0,23	0,29	0,38	0,45	0,54	1,31						
avr-16	chiffre d'affaire en k€	non-renseigné	< 0	0-57	57-92	92-119	119-144	144-169	169-222	222-386	386-2750	2750-7936	7937 et +						
	coefficient	0,00	-1,10	-1,33	-0,36	0,04	0,20	0,28	0,34	0,48	0,57	0,66	1,33						
juil-16	chiffre d'affaire en k€	non-renseigné	< 0	0-59	59-93	93-120	120-145	145-170	170-223	223-291	291-454	454-660	660-830	830-1594	1594-2770	2770-7976	7978 et +		
	coefficient	0,00	-1,47	-1,72	-0,71	-0,27	-0,07	0,03	0,13	0,28	0,37	0,46	0,57	0,64	0,80	1,04	1,69		
oct-16	chiffre d'affaire en k€	non-renseigné	< 0	0-60	60-94	94-121	121-146	146-171	171-197	197-225	225-293	293-391	391-458	458-666	666-839	839-1611	1612-2805	2805-8096	8097 et +
	coefficient	0,00	-1,56	-1,80	-0,78	-0,33	-0,11	0,01	0,10	0,18	0,30	0,35	0,43	0,48	0,61	0,67	0,81	1,06	1,63
janv-17	chiffre d'affaire en k€	non-renseigné	< 0	0-52	52-80	80-107	107-132	132-156	156-207	207-237	237-272	272-509	509-622	622-1040	1040 et +				
	coefficient	0,00	-1,55	-1,89	-0,88	-0,43	-0,03	0,16	0,38	0,61	0,72	0,84	0,95	1,06	1,17				

Tableau 2 – Estimation des paramètres associés aux classes des chiffres d'affaires (secteur de la construction)

2.6. Les nouveaux modèles corrigent-t-il totalement le biais de non-déclaration ?

L'objectif principal de toutes les méthodes de redressement est de réduire le biais de non-réponse, qui est dû au fait que les déclarants présentent des caractéristiques différentes des non-déclarants. La correction et l'ampleur du biais de non-réponse dépend du type de non-réponse présent dans les données :

- un mécanisme de non-réponse est uniforme (« Missing Completely At Random ») si la probabilité de non-réponse ne dépend pas des variables d'intérêt ($\forall i,\ p_i = p$) . Ce type de mécanisme au niveau de la population n'est généralement pas réaliste, mais s'avère particulièrement efficace dans notre contexte de classes de repondération (cf. *supra*) ;
- un mécanisme de non-réponse est ignorable (« Missing At Random »), si après avoir pris en compte toute l'information auxiliaire appropriée *x,* la probabilité de réponse ne dépend pas des variables d'intérêt ($P(r_i = 1/y, x) = P(r_i = 1\ /\ x)$).
- si un mécanisme de non-réponse n'est pas ignorable, il est qualifié de non-ignorable (« Non Missing At Random » ; lorsqu'il existe un lien de causalité entre la variable d'intérêt et la probabilité de non-réponse). Dans ce cas, les statistiques sur la variable d'intérêt sont biaisées.

Il est difficile de connaître le type de mécanisme de non-réponse qui se trouve dans les données, et de déterminer l'ampleur du biais de non-réponse induit. Toutefois, introduire les DPAE (Déclarations préalables à l'embauche) dans les modèles fournit des informations (limitées) sur le biais.

S'il s'avère difficile de mesurer l'ampleur d'un éventuel biais dû à la non-déclaration, il est néanmoins possible de juger de l'ampleur de l'erreur totale, et par conséquent d'évaluer l'efficacité de la correction apportée. Les estimations obtenues à partir de l'échantillon de déclarants peuvent, en effet, être comparées aux agrégats connus, qui sont eux-mêmes calculés à partir des bases de référence.

2.6.1. La Déclaration préalable à l'embauche (DPAE)

Généralement, il est impossible de savoir si l'on se trouve dans une situation d'un mécanisme de non-réponse ignorable, ou d'un mécanisme de non-réponse non-ignorable. Néanmoins, un taux d'entrée dans les établissements est calculable *via* les DPAE. Cette variable est *a priori* une bonne approximation du taux d'entrée issu des DSN. Par conséquent, en testant le pouvoir explicatif de ces variables dans les modélisations de la non-déclaration, il est possible d'avoir une idée du mécanisme de non-réponse :

- L'hypothèse nulle des tests de nullité appliqués aux paramètres associés aux taux d'entrée en CDI dans les établissements entre les mois de mai 2015 et décembre 2016 est acceptée dans plus de 80 % des cas au niveau 1 % (cf. tableau 3) ;
- Quant aux tests de nullité des paramètres associés aux taux d'entrée en CDD, l'hypothèse nulle est acceptée dans un peu plus de 70 % des cas.

Ainsi, le comportement de réponse à la DSN ne dépendrait que faiblement des flux de main-d'œuvre des établissements. Cette absence d'effet de causalité signifierait que l'on se trouve dans une situation d'un mécanisme de non-réponse ignorable, c'est-à-dire une situation où le biais de non-déclaration est corrigé par le traitement par pondération décrit précédemment.

	construction	industrie	tertiaire	Ensemble
taux d'entrée en CDD	75	70	70	**72**
taux d'entrée en CDI	85	95	70	**83**

Tableau 3 – Part des mois entre mai 2015 et décembre 2016 pour lesquels la partie non-expliquée des modèles n'est pas corrélée avec les DPAE

Note de lecture : les variables sont issues des DPAE ; les tests ont été appliqués pour chaque mois entre mai 2015 et décembre 2016 ; part (en %) de test dont l'hypothèse nulle a été acceptée au niveau 1 %.

2.6.2. L'erreur totale

S'il n'est pas possible de calculer le biais causé par la non-déclaration de certains établissements, il est toutefois possible de connaître l'erreur totale, c'est-à-dire l'erreur résiduelle après redressement des données. Pour cela, on compare les estimations obtenues à partir de l'échantillon de déclarants (notées $\widehat{X_\pi}$), aux agrégats connus calculés sur les bases de référence (répondants et non-répondants confondus ; notés *X*). L'erreur totale s'écrit alors sous la forme $\theta = \widehat{X_\pi} - X = \sum_{i\,\epsilon\,R} p_i . X_i - \sum_{i\,\epsilon\,bdr} X_i$ et l'erreur relative $\varepsilon = \frac{(\widehat{X_\pi} - X)}{X}$. Pour que le calcul soit rendu possible, il est nécessaire que la variable soit renseignée pour tous les établissements des bases de référence. En conséquence, trois variables sont retenues :

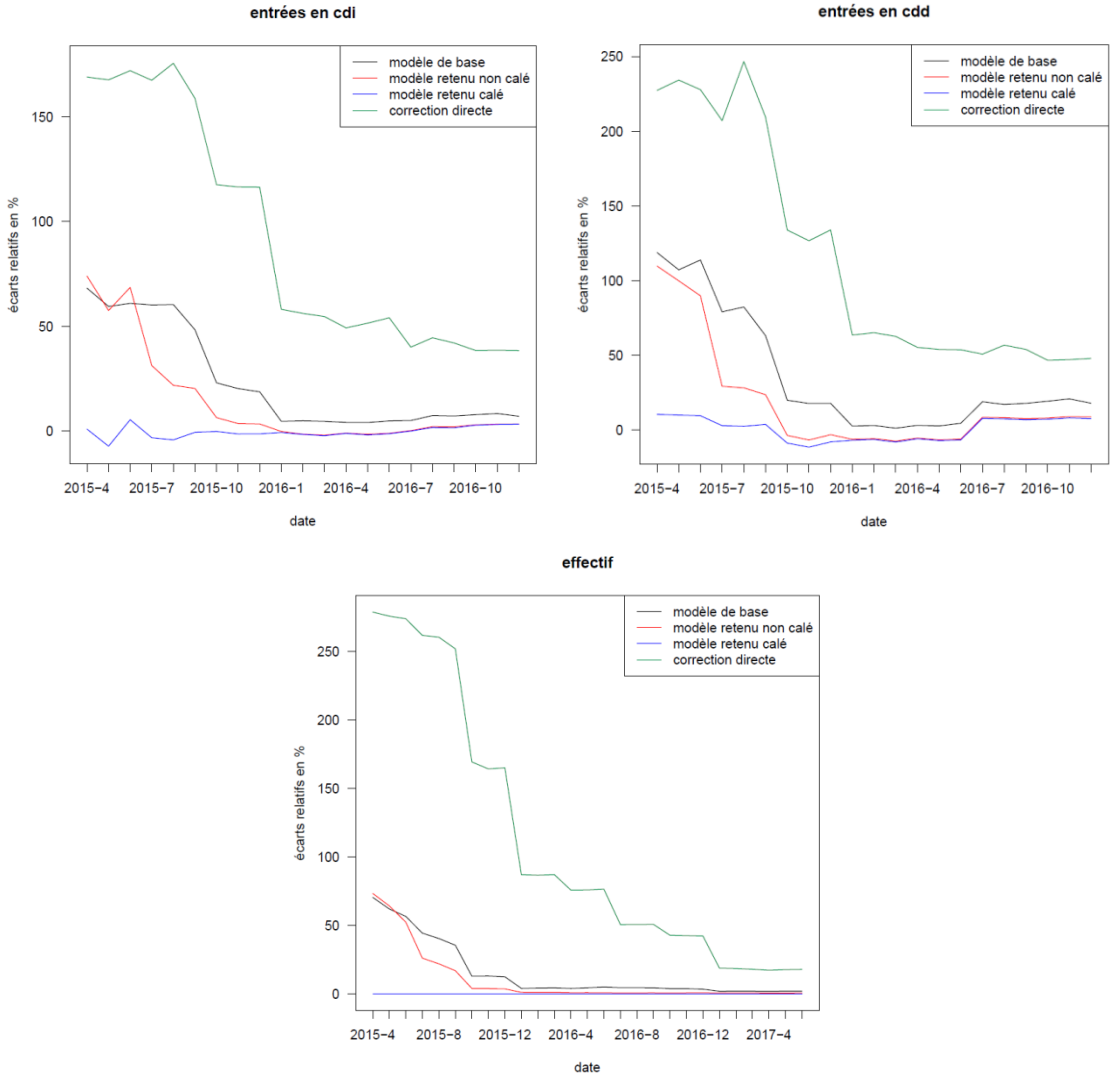
- l'effectif de référence ;
- le nombre d'entrées en CDI (issu des DPAE) ;
- le nombre d'entrées en CDD (des DPAE).

En outre, l'obtention des estimations nécessite l'utilisation de jeux de pondérations. Ici, différentes estimations sont confrontées. Elles sont calculées à partir :

- des poids constants (égaux à l'inverse du taux de déclaration ; « correction directe » efficace et suffisante dans le cas d'un mécanisme de non-réponse uniforme) ;
- des poids non calés issus du modèle « de base », après création des groupes de réponse homogène (GRH) ; « correction modèle de base » ;
- des poids non calés issus des modèles retenus, après création des GRH ; correction « modèles retenus » ;
- des poids calés issus des modèles retenus (calage non présenté dans ce document).

Globalement, les estimations obtenues *via* les poids issus des modèles retenus minimisent l'erreur totale, comparativement à celles qui sont obtenues avec les poids du modèle de base (cf. graphiques 14 à 16). Plus particulièrement, si la correction « modèle de base » surestime, en moyenne, de 18,7 % le nombre d'entrées en CDD (respectivement 7,1 % en CDI) au second semestre 2016, la correction « modèles retenus » le surestime de 8,4 % (resp. 2,3 %). À noter qu'en utilisant les poids calés issus des modèles retenus, la surestimation est réduite à 7,4 % pour le nombre d'entrées en CDD, et s'établit à 2,5 % pour le nombre d'entrée en CDI.

Ces erreurs ne sont pas forcément le signe d'un biais de non-déclaration, et peuvent être le reflet de la variance des estimateurs. Aussi, les estimations, calculées avec les pondérations issus des modèles retenus (calées ou non), sont inférieures au nombre d'entrées relevées par les DPAE au premier semestre 2016.



Graphiques 14, 15 et 16 – Comparaison des estimations effectuées à partir des déclarants, avec les grandeurs connues dans les bases de référence

Conclusion

Les statistiques des mouvements de main-d'œuvre ne sont désormais plus construites à partir de la déclaration des mouvements de main-d'œuvre et de son enquête associée (DMMO/EMMO), mais sont reconstituées à partir des déclarations sociales nominatives (DSN). Une méthodologie importante a été mise en place à la Dares pour assurer ce changement de source. En particulier, un redressement de type « non-réponse totale » était nécessaire, en raison de la montée en charge progressive du dispositif des DSN, afin de pallier la non-déclaration de certains établissements et de garantir la représentativité des données finales.

La méthodologie adoptée comprend différentes étapes. La probabilité de déclaration des établissements est tout d'abord estimée. Puis des classes de pondérations sont construites, à partir de ces probabilités estimées, par la méthode d'Haziza Beaumont (2007) ; les pondérations affectées aux établissements déclarants correspondent à l'inverse du taux de réponse dans chacune des classes. Enfin, un calage sur des marges d'effectif modifie les pondérations obtenues pour minimiser la variance des estimateurs finaux.

Ce présent document restitue essentiellement les modélisations retenues pour estimer les probabilités des établissements à déposer des DSN. Des modélisations distinctes selon les trois grands secteurs (industrie, construction, tertiaire) ont été utilisées. Elles intègrent plusieurs variables, comme le secteur détaillé ou la région d'activité. Par ailleurs, les variables quantitatives, telles que l'effectif de référence de l'établissement ou le chiffre d'affaire de l'entreprise associée sont transformées en variable qualitatives. Cette catégorisation permet notamment de prendre en compte les non-linéarités existantes dans les relations entre ces variables et la propension d'un établissement à effectuer une DSN. Si les spécifications des modèles diffèrent d'un secteur à l'autre, elles diffèrent également dans le temps. La catégorisation des variables quantitatives est en effet adaptée à chaque phase de montée en charge de la DSN.

Les données de la déclaration préalable à l'embauche, gérées par l'Acoss, sont particulièrement corrélées aux variables d'intérêt de mouvement de main-d'œuvre. Connues à la fois pour les établissements déclarants et non-déclarants en DSN, elles s'avèrent être une variable auxiliaire intéressante pour vérifier la qualité de la correction apportée. D'une part, leur introduction dans les modèles retenus ne diminue pas les parties inexpliquées de ces derniers, laissant ainsi penser que le processus de non-déclaration est ignorable (*i.e.* : les statistiques finales ne sont pas biaisées par la non-déclaration). D'autre part, le calcul des erreurs totales est possible à partir de ces données : l'ampleur des erreurs semble modérée, surtout lorsque l'on prend en compte dans la modélisation les non-linéarités dans la relation entre l'effectif d'un établissement (ou le chiffre d'affaires de l'entreprise) et la probabilité de l'établissement à remettre des DSN.

Bibliographie

- Ardilly P. (2006), « Les techniques de sondage », *édition Technip*, juin
- Barlet M., Raynaud P., Sanzéri O. (2015), « Les statistiques sur les mouvements de main-d'œuvre : passage d'une enquête à une source administrative », *Journées de méthodologie statistique*, mars
- Da Silva D.N., Opsomer J.D. (2009), « Nonparametric Propensity Weighting for Survey Nonresponse Through Local Polynomial Regression », *Survey Methodology*
- Haziza, D., Beaumont, J.-F. (2007), « On the construction of imputation classes in surveys », *International Statistical Review*, mars
- Milin K. (2018), « Reconstitution des mouvements de main-d'œuvre à partir des déclarations sociales nominatives : guide méthodologique », *Document d'études*, Dares, n° 221, juin.